知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【专稿】

产业链视角下结合 K-means 和 LDA 的专利技术主题挖掘与趋势分析

——以虚拟现实技术为例

陈玲[1] 林平[1] 段尧清[1,2]
[1] 华中师范大学信息管理学院 武汉 430079
[2] 湖北省数据治理与智能决策研究中心 武汉 430079

摘要：[**目的 / 意义**] 在产业链视角下，以虚拟现实技术为例，构建 VR 专利产业链语料库，挖掘中国 VR 专利的技术主题、研发热点和未来发展趋势。[**方法 / 过程**] 首先，利用 Python 爬取 VR 领域的专利文本，通过数据清洗得到有效语料库；然后，结合 IPC 分类号和 K-means 聚类算法，构建并验证 VR 专利产业链；最后，基于 TF-IDF 算法和 LDA 主题模型，识别出产业链视角下中国 VR 专利的核心技术主题及其综合强度、技术研发热点和未来趋势。[**结果 / 结论**] 当前中国 VR 产业链各环节的专利比例不均衡，上游研发最热门，其次是下游应用，最薄弱的是中游制作。主题挖掘方面，上游热点为软件研发，中游热点为影视制作，下游热点为医疗、教育、娱乐应用。未来趋势方面，产业链上游将以电数字数据处理、光学元件、图像通信等技术为主流，中游将以车辆部件、动力装置、减振装置等技术为主流，下游将以室内游戏、医学诊断、鉴定等技术为主流。

关键词：K-means 聚类算法　LDA 主题模型　技术主题演化　文本挖掘　VR（虚拟现实）
分类号：G250
DOI：10.13266/j.issn.2095-5472.2020.013

引用格式：陈玲 , 林平 , 段尧清 . 产业链视角下结合 K-means 和 LDA 的专利技术主题挖掘与趋势分析——以虚拟现实技术为例 [J/OL]. 知识管理论坛 , 2020, 5(3): 135-146[引用日期]. http://www.kmf.ac.cn/p/208/.

1 引言

专利是衡量科学技术发展的重要指标，专利内容挖掘是提高科学技术竞争力的主要途径之一。专利内容挖掘涉及专利分类、专利聚类、主题识别、技术趋势分析等方面，其中专利技术主题分析是其研究的核心所在。专利技术主题分析聚焦于识别专利文本的主题（如对主题

基金项目：本文系国家社会科学基金重点项目“基于全生命周期的政府开放数据整合利用机制与模式研究”（项目编号：17ATQ006）研究成果之一。
作者简介：陈玲 (ORCID：0000-0003-0379-3512)，博士研究生，E-mail：2471685835@qq.com；林平 (ORCID：0000-0003-0283-6824)，硕士研究生；段尧清 (ORCID：0000-0002-8991-5842)，教授，博士生导师。
收稿日期：2020-03-20　发表日期：2020-06-05　本文责任编辑：刘远颖



知识管理论坛
2020 年 第 3 期(总第 27 期)

进行分类、构建主题间的相互关系、预测主题的发展趋势等），对技术研发内容具有高度的概括性和代表性[1]。随着深度学习和机器学习的兴起，文本挖掘被越来越广泛地应用在技术专利主题分析中，其中以 LDA（Latent Dirichlet Allocation）主题模型尤为突出。专利技术主题分析方法主要是抽取专利文献标题、摘要及技术要点中的技术特征词，利用文本挖掘方法选择获得主题词，建立主题词之间的共现关联关系，从而聚类获得技术主题[2]。专利技术主题分析常用的方法包括：①利用专利的分类属性作为其技术主题；②通过专利共现网络和引用关系为专利聚类；③使用 SAO（subject -action -object）结构语义相似度识别、主题模型或主题聚类等方式从专利等科技文献中挖掘技术主题；④借助技术主题的时间信息，使用时间序列分析等方式预测技术主题演化趋势[3]。

在信息技术快速发展的知识经济时代，虚拟现实作为战略新兴技术的代表，涉及通信、互联网、新媒体等多个领域，具有突出的跨界融合性与技术交叉性，有望引领新一轮技术的变革。众多科技新兴企业均在 VR 领域积极布局，主要科技大国也均把 VR 列为战略新兴领域，中国在国家“十三五”规划纲要、G20 工商峰会上的重要讲话中提出要发展人工智能和虚拟现实等技术，大力支持虚拟现实（VR）等新兴前沿领域创新和产业化，建设创新型世界经济[4-8]。在产业链视角下，深度挖掘中国 VR 领域的专利技术主题、技术热点与发展趋势，可以分别从宏观、中观和微观不同的角度对政府、产业和企业提供不同的情报服务，在此基础上制定相应的竞争战略；有助于相关政府部门、VR 科研机构和企业等主体在中国和全球范围内更好地进行专利布局，为中国 VR 产业发展提供参考建议，最终提高中国 VR 领域的整体产业竞争力。

2 相关研究

2.1 虚拟现实

虚拟现实是以计算机技术为核心，生成与现实环境在视、听、触感等方面高度近似的数字化环境。用户借助相关设备与虚拟环境中的对象进行交互，从而产生真实环境的感受和体验。目前关于虚拟现实的研究主要集中在技术研究[4-5]、系统研究[6-7]、应用研究[8]3 个方面：①虚拟现实技术研究。学者主要从立体显示技术[9]、传感器技术[10]、三维图形生成技术[11]等方面将虚拟和现实环境进行混合、实时交互、三维注册。②虚拟现实系统研究。主要分为硬件研究和软件研究，硬件研究包括三维跟踪定位设备、人体运动捕捉设备、触觉力觉反馈设备等的研究[12]；软件研究包括数据库研究[13]，三维动画、网络场景等应用软件研究[14]，基于 Vizard 软件、Virtools 软件、EON 软件等的虚拟现实开发平台研究[15]。③虚拟现实应用研究。随着技术不断地进步与成熟，虚拟现实技术逐渐被应用到教育[16]、医疗[17]、图书馆[18]、博物馆[19]等不同场合，从而为人们的生产、生活、学习带来巨大的影响与冲击。

2.2 基于文本挖掘的专利技术主题分析

技术主题分析是文本挖掘在专利分析中的重要应用之一。目前已有较多利用文本挖掘方法进行专利技术主题分析的研究成果，依次包括词频统计分析、共词分析、文本聚类分析、文本挖掘技术与引文聚类相结合的技术主题分析[20]。①基于词频统计的技术主题研究。主要是通过 IPC 分类号、高频词等的统计分析，研究某技术领域的主题分布情况[21]。②基于共词分析的技术主题研究。主要包括共词网络分析、共词聚类分析和战略图分析 3 种方法，可以比较客观地揭示技术领域中的各技术主题及技术主题之间的相互关联[22-23]。③基于文本聚类的技术主题研究。主要是对专利进行聚类，形成代表技术主题的多个聚簇；为每个聚簇生成主题词，从而直观有效地表示技术主题的分布情况[24]。④基于文本挖掘与引文聚类相结合的技术主题研究。主要从文本信息与引用信息的底层融合角度，分析技术研究热点、识别新兴技术主题、预测技术主题的发展趋势[25]。

知识管理论坛，2020(3):135-146
DOI：10.13266/j.issn.2095-5472.2020.013

2.3 基于 LDA 主题模型的专利技术主题分析

基于 LDA 模型的专利技术主题分析主要分为两类，一类是直接采用传统的 LDA 模型分析专利文本构成的语料库，如对专利领域技术信息进行主题划分、测量与分析专利丛林 [26]，挖掘专利领域的技术及其继承关系 [27]。另一类是根据特定的分析目的或专利信息的结构特征对 LDA 模型进行改进或拓展，如构建基于 SAO 结构、P&S 模式的 LDA 主题模型 [1]，提出结合 LDA 和 HMM 的组合方法 [3]，构建基于 IPC 和 WI 结构的 WI-LDA 模型 [28] 等，分析某一专利领域的技术主题分布，识别和预测专利领域的核心技术、演化规律及未来趋势。

2.4 文献述评

已有文献中，关于专利技术主题的分析，多将专利文本视为统一整体进行文本挖掘，或者按照专利标题、专利关键词、专利正文等不同视角，进行主题挖掘，较少结合专利的产业链特性进行技术主题分析。而关于产业链视角下的相关专利分析，在产业链构建方面均是通过经验判断等定性研究方法进行专利挖掘，且多从专利数量、专利类型、地域分布、核心申请主体等角度出发，进行专利分布研究，未结合专利文本进行技术主题的深度挖掘。在此背景下，本研究从产业链视角出发，以 VR 技术为例，利用专利的 IPC 分类号构建专利产业链，并通过 K-means 聚类验证产业链，能够为专利领域产业链研究提供新的研究视角；基于产业链语料库，采用 TF-IDF 算法和 LDA 主题模型，深度挖掘中国 VR 领域的专利技术主题、技术热点与发展趋势，能够为专利领域技术研究提供新的研究思路、为 VR 领域扩展研究内容。

3 研究设计

3.1 研究框架

以中国 VR 专利的相关数据为原始语料库，在归并处理、噪音清除、加工分组等数据清洗的基础上，得到有效专利语料库。基于有效语料库，进行文献调研和专家咨询，利用专利的 IPC 分类号和 K-means 聚类算法，构建并验证 VR 专利产业链，得到基于 IPC 编码和聚类的产业链语料库。基于产业链语料库，进行文本分析，利用 TF-IDF 算法计算关键词权重，利用 LDA 主题模型挖掘各环节的技术主题及主题词权重，识别产业链视角下的核心技术主题及其主题强度，分析中国 VR 专利的技术研发热点和未来趋势。具体研究框架如图 1 所示：

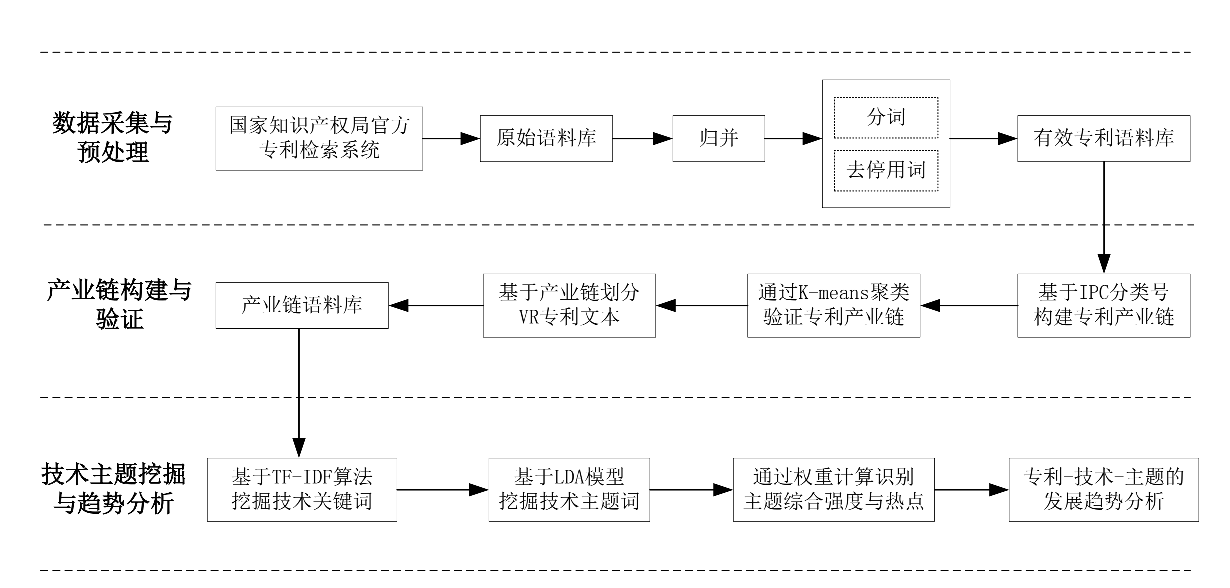


图 1 研究框架



知识管理论坛
2020 年 第 3 期(总第 27 期)

3.2 数据采集

本文的目标数据库确定为国家知识产权局的官方专利检索系统，检索式设定为：发明名称 =(虚拟现实 OR VR OR virtual reality) OR 摘要 =(虚拟现实 OR VR OR virtual reality) OR 关键词 =(虚拟现实 OR VR OR virtual reality)，检索时间为 2019 年 5 月 10 日，在过滤条件中勾选"有效专利"复选框，获得有效专利 14 372 件。目前中国常用的专利信息源包括：中华人民共和国国家知识产权局的官方专利检索系统、国家知识产权出版社主办的中国知识产权网的专利检索系统、中国专利信息网专利检索系统、北京市经济信息中心易信网的专利检索系统等[29]。其中，国家知识产权局的官方专利检索系统是检索中国专利的官方网站，最具有权威性，其数据收录主体范围涵盖广泛，收录数据信息类别全面，数据更新频率较高且更新时间具有周期性和规范性。

3.3 数据清洗

对检索得到的专利数据进行去重、筛选和加工，简要说明如下：①归并处理。对同一专利权人在不同的专利记录中可能登记有不完全相同的名称，进行归并处理。②噪音清除。阅读并删除与所检主题不相关的专利数据，对语料库依次进行大小写转换、去标点、去数词等去噪处理。③加工分组。根据专利固有格式与领域特点，对采集的数据进行加工、分组，建立符合研究需要的专题子数据库[30]。最终确定 12 380 件专利数据用于构建 VR（虚拟现实）领域有效专利语料库。

其中，去噪是尤为关键的环节，主要包括分词和去停用词两部分。据此，本文将专利摘要内容整合在 TXT 文档中作为文本信息，在 Python 语言环境下，对每一条摘要数据进行分词、去除停用词。分词使用 Python 中的专业分词模块 jieba，选择精准模式将句子尽可能精确切分，并将"外观设计""实用新型""发明专利"等具有代表性意义的词组添加到自定义词典，避免关键词汇被拆分，影响后续文本分析。分词完成后利用停用词表将分词后的数据进行进一步的清洗，过滤分词结果中的噪声。自定义的主要停用词包括"虚拟现实""VR"（因本文研究虚拟现实领域专利，为避免"虚拟现实""VR"出现频率过高影响其他高频词的凸显，故将其停用）、"所述""提供""包括""用于""省略""涉及""获取""建立""选择""要点""特征""连接""之间""步骤"等，表 1 随机列举了 4 条专利摘要原文及其对应的分词结果。

表 1 随机列举 4 条专利摘要原文及其对应的分词结果

专利摘要原文	分词结果
在灯罩未与灯杆叠合时,做台灯使用,叠合后灯罩的透明窗（A）部位又可做夜灯使用	灯罩 未 灯杆 叠合 时 做 台灯 叠合 灯罩 透明 窗 部位 做 夜灯
本外观设计产品是一种卡通玩偶台灯,在灯罩未与灯杆叠合时,做台灯使用；叠合后灯罩的透明窗又可做夜灯使用	外观设计 产品 卡通 玩偶 台灯 灯罩 未 灯杆 叠合 时 做 台灯 叠合 灯罩 透明 窗 做 夜灯
虚拟现实头盔的前端与图像显示眼镜连接	头盔 前端 图像 显示 眼镜 连接
本实用新型涉及一种视点平移的虚拟空间实景视频图像生成装置,属于多媒体虚拟现实技术领域	实用新型 涉及 视点 平移 虚拟空间 实景 视频 图像 生成 装置 多媒体 虚拟现实 技术 领域

3.4 研究指标

专利情报分析是在对专利情报进行筛选、整理的基础上，利用统计方法和手段，对其中所含的各种情报要素进行统计、排序、对比、分析和研究，从而了解技术发展的过去和现状。通常来说，专利情报分析主要有两种：定量分析和定性分析。定量分析是指对专利文献的外部特征按照指定的指标进行统计，再对收集到的数据进行解释和分析；定性分析则是通过对专利的内容进行技术归纳，得出有效的分类和结论[31]。本文的专利研究指标及其作用具体如表 2 所示：

知识管理论坛，2020(3):135–146
DOI：10.13266/j.issn.2095–5472.2020.013

表 2 产业链视角下的专利指标分析表

专利分析	类型	专利分析指标	专利分析目的
技术主题挖掘与趋势分析	定性+定量	技术关键词挖掘	识别产业链各环节的技术关键词
		技术主题词挖掘	掌握该行业和竞争对手的技术研发侧重点
		技术主题强度分析	通过综合主题强度和各主题的强度分布，识别研发热点
		技术发展趋势分析	识别产业链各环节的专利-主题-技术领域的未来趋势

4 结合 IPC 分类号和 K-means 算法的 VR 专利产业链分析

已有文献在构建产业链方面均是基于人为定义、解读，划分上中下游各个环节。本文则是将 VR 相关的全部专利检索获得后进行数据清洗（保证了专利产业链的检全率），基于 IPC 分类号划分上中下游，并基于 K-means 算法进行上中下游的二次验证（保证了专利产业链的准确率）。

4.1 基于 IPC 分类号的专利产业链构建

依据文献调研、专家咨询和专利的 IPC 分类号，将虚拟现实产业链分为工具 / 设备设计、内容制作、行业应用。在此基础上，选取专利的 IPC 分类号作为语义情景的限定，为所有专利赋予产业链语义。在提取 IPC 分类号时，不同的 IPC 层级会产生不同的聚类效果。基于 IPC 大类的划分过于粗泛，聚类效果不明显；基于 IPC 大组的划分过于密集，同样不适合聚类；而基于 IPC 小类的划分，能够在区分度明显的基础上保证规模不过于巨大，因而最终选定以主 IPC 分类号小类作为语言情景的限定。为了研究过程的简易性及结果展示的直观性，将专利数据涉及的产业链与 IPC 小类进行编码，部分编码分布情况如表 3 所示。其中，产业链上游为“工具 / 设备设计”，产业链中游为“内容制作”，产业链下游为“行业应用”。

表 3 中国虚拟现实领域产业链与 IPC 小编码分布情况（部分）

产业端	产业链	产业链编码	IPC 小类	技术领域含义	小类编码
上游	工具/设备设计	01	G06K	数据识别；数据表示；记录载体；记录载体的处理	152
			G02B	光学元件、系统或仪器	141
			H04J	多路复用通信	194
			H05K	电设备的外壳或结构零部件；电气元件组件的制造	204
中游	内容制作	02	A47C	椅子	15
			B43L	书写或绘图用品；书写或绘图辅助用品	53
			C12C	啤酒的酿造	83
			D03D	机织织物；织造方法；织机	86
			F16P	一般安全装置	115
下游	行业应用	03	A61C	牙科；口腔或牙齿卫生的装置或方法	20
			A63B	体育锻炼、体操、游泳、爬山或击剑用的器械；球类；训练器械	30
			E04B	一般建筑物构造；墙，例如：间壁墙；屋顶；楼板；顶棚；建筑物的隔绝或其他防护	94



4.2 基于关键词聚类的专利产业链验证

研究采用 K-means 算法验证已构建的专利产业链。首先，合并“同类关键词”。通过人工观察，将包含“本发明”“本发明专利”“本专利”等数据的关键词，统一合并为“发明专利”。其次，采用 K-means 算法中的欧氏距离来计算数据对象间的距离。根据相似性原则，将具有较高相似度的数据对象划分至同一类簇，将具有较高相异度的数据对象划分至不同类簇。

VR 专利摘要文本的 K-means 聚类效果如图 2 所示。依据产业链的分类特性和已有关于产业链划分的研究文献可知，专利产业链通常划分为上、中、下 3 类 [32] 或基础、技术、应用 3 类 [33-34]。据此，研究将类簇个数 K 值设定为 3，将专利文本聚集成 3 类主题。从图 2 中可以看到 3 个类簇有效地分隔开来，相似主题的文献聚集在一起，文本聚类效果较好。其中黄色表示“工具 / 设备设计”主题，紫色表示“行业应用”主题，绿色表示“内容制作”主题；且“工具 / 设备设计”专利聚类数量＞“行业应用”专利聚类数量＞“内容制作”专利聚类数量。观察聚类结果可知，“工具”“设备”等关键词聚为一类，划分至“工具 / 设备设计”专利类别；“游戏生产”“声音生产”“视频生产”等关键词聚为一类，划分至“内容制作”专利类别；“医疗应用”“教育应用”“旅游应用”等关键词聚为一类，划分至“行业应用”专利类别。基于关键词聚类的 VR 专利产业链验证结果，与上文中基于 IPC 分类号的 VR 专利产业链构建结果具有一致性。据此，根据产业链的构建和验证结果，对中国 VR 专利进行分类，构建产业链语料库。

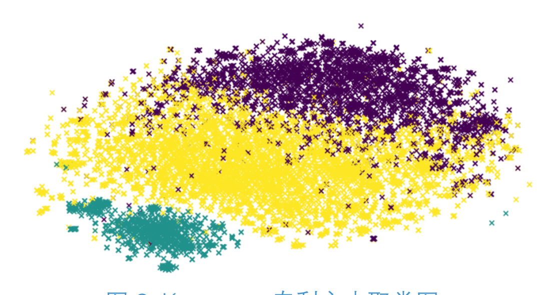


图 2 K-means 专利文本聚类图

5 产业链视角下中国 VR 专利的技术主题与趋势分析

5.1 基于 TF-IDF 算法的技术关键词挖掘

为了避免 LDA 主题分析抽取出的特征词汇不具主题代表性，研究首先使用 TF-IDF 算法对所得词汇赋予不同权重，有效过滤常见词汇，保留重要词汇，进而提高主题特征词的抽取准确率。TF-IDF 是一种计算词语权重的经典统计方法，由词频 (term frequency，TF) 和逆向文档频率 (inverse document frequency，IDF) 两部分数据组成。TF-IDF 的计算如公式（1）所示，其中，$tf_{i,j}$ 代表词语 w_i 在文档 d_j 中出现频率，idf_i 代表词语 w_i 在文本库 d 中的逆向文档频率。通过公式可以看出，词语 w_i 对文档 d_j 的重要程度和它在文档 d_j 中出现的频率成正比，和它在整个文本库 d_j 中包含词语 w_i 的文档数成反比。

$$tf-idf_{i,j}=\frac{tf_{i,j}\times idf_i}{\sqrt{\sum_{w_i\in d_j}[tf_{i,j}\times idf_i]^2}} \quad \text{公式（1）}$$

依照产业链语料库数据和编码分词，在 Python 中提取摘要文本关键词；整合相似的文本数据，删除无实际作用的字段，根据 TF-IDF 算法计算关键词权重。TF-IDF 算法是通过计算特征词在整个文本库中出现的总频率，从而标记出关键词的重要程度。产业链各环节中国 VR 专利摘要文本的高频关键词及权重计算结果如表 4 所示，可以看出“发明专利”类型在产业链上、中、下游的比重均较大。此外，产业链上游“工具 / 设备设计”中“发明专利”类型占比较大，产业链中游“内容制作”中“外观设计”类型专利占比较大，产业链下游“行业应用”中“实用新型”类型专利占比较大。

5.2 基于 LDA 模型的技术主题词挖掘

在基于 LDA 模型的主题挖掘中，最佳主题数目的确定是最为关键的一步，本文使用 Gibbs 采样的方法推断 LDA 模型中所涉及的多个分布。首先，充分参考虚拟现实产业链环节数量后，将各个环节的输出主题数目初步确定为 3-10 个，对 LDA 模型进行训练。其次，通过计算模型困

知识管理论坛，2020(3):135–146
DOI：10.13266/j.issn.2095–5472.2020.013

惑度 Perplexity 来判断模型的好坏，从而确定该模型的最佳参数，即使用不同数量的主题分别建模，随机将语料库划分为训练集与测试集，训练集和测试集比例为 8:2。最后，通过计算困惑度 10 次结果的平均值将产业链上游的最佳主题数目确定为 4 个，将产业链中游的最佳主题数目确定为 3 个，将产业链下游的最佳主题数目确定为 7 个。

表 4 产业链各环节专利摘要高频关键词及其权重

上游关键词	TF-IDF权重	中游关键词	TF-IDF权重	下游关键词	TF-IDF权重
发明专利	0.433 685 602	外观设计	0.154 538 305	实用新型	0.146 471 830
视图	0.298 644 789	模块	0.105 986 349	设备	0.101 432 055
装置	0.193 443 350	图像	0.096 887 133	发明专利	0.096 677 809
设备	0.104 926 572	场景	0.082 635 769	模块	0.093 789 460
工具	0.090 669 127	系统	0.076 838 141	连接	0.089 111 996
灯罩	0.090 018 305	用户	0.072 846 204	眼镜	0.085 453 807
俯视图	0.083 199 331	发明专利	0.069 709 442	用户	0.085 028 156
连接	0.081 526 927	内容	0.065 064 784	图像	0.073 168 758
游戏	0.081 252 223	装置	0.064 204 349	装置	0.073 157 422
仰视	0.079 616 895	显示	0.060 203 497	场景	0.070 251 633
软件系统	0.075 799 232	仿真	0.059 405 618	设置	0.064 731 162
显示器	0.074 215 894	现实	0.057 281 859	固定	0.052 306 909

通过充分了解该领域的技术知识，对中国虚拟现实领域专利进行技术主题标注，确定主题名称。某种程度上，使用 LDA 主题模型挖掘到的主题可视为从技术链角度对虚拟现实技术进行细分，如表 5 所示。由表 5 可以看出，每个主题之间的区分非常明显。产业链上游——“工具 / 设备设计”的 4 个主题分别为输入设备、显示设备、拍摄设备、软件；产业链中游——“内容制作”的 3 个主题分别为影视、声音、游戏；产业链下游——“行业应用”的 7 个主题分别为房地产、旅游、工业、军事、医疗、教育、娱乐。

5.3 产业链视角下的技术主题强度与热点分析

经文献研究与小组讨论认为，产业链视角下技术主题强度的衡量指标主要包括：产业链各环节的专利数量权重与专利文档概率。其中，产业链各环节的专利权重为上、中、下游专利数量在总专利数量中的占比，文档概率为上、中、下游产业链视角下各主题的隶属概率值。产业链视角下各技术主题的综合强度计算如公式(2)所示：

$$TI_i = \left(LDA_{i^*} \frac{n_{ip}}{\sum_{p=1}^{3} n_{ip}} \right) \Bigg/ \left(\sum_{i=1}^{15} LDA_{i^*} \frac{n_{ip}}{\sum_{p=1}^{3} n_{ip}} \right)$$

公式（2）

TI_i 为第 i 个主题的综合强度。其中，i 为专利的 15 个主题（i=1，2，3，……，14），p 为这 14 个主题分别对应的三个产业链环节（p=1，2，3）。LDA_i 为第 i 个主题的 LDA 权重值，n_{ip} 为第 i 个主题所对应的第 p 个产业链环节的专利数量，$\frac{n_{ip}}{\sum_{p=1}^{3} n_{ip}}$为第 p 个产业链环节的专利数量权重值；$LDA_i * \frac{n_{ip}}{\sum_{p=1}^{3} n_{ip}}$为第 i 个主题的 LDA 权重值与其所对应的第 p 个产业链环节的专利数量权重值之乘积。

根据 LDA 模型提取的权重值，结合产业链各环节的专利数量，计算出产业链视角下各技术主题的综合强度分布，结果如表 6 所

表 5 LDA 模型主题词识别结果

产业链		主题	主题词（隶属概率值）
上游设计	工具/设备设计	输入设备	输入（0.072）、主机（0.062）、游戏（0.037）、显示器（0.032）、穿戴（0.029）、应用（0.028）、姿态（0.028）、头部（0.027）……
		显示设备	图像（0.015）、显示（0.014）、模块（0.014）、数据（0.011）、信息（0.012）、用户（0.011）、场景（0.013）、设备（0.015）……
		拍摄设备	对象（0.067）、交互（0.041）、配置（0.036）、影像（0.021）、视频（0.019）、图像处理（0.017）……
		软件	通信（0.072）、光学（0.123）、处理器（0.106）、本体（0.047）、存储（0.044）、识别（0.024）、无线连接（0.019）……
中游生产	内容制作	影视	电影（0.036）、装置（0.031）、场景（0.029）、图像（0.028）、显示（0.026）、影视（0.025）、屏幕（0.023）……
		声音	拍摄（0.027）、声音（0.017）、物体（0.016）、传感器（0.014）……
		游戏	外观设计（0.063）、游戏（0.013）、眼镜（0.028）、手柄（0.026）、体验（0.014）、立体图（0.056）……
下游应用	行业应用	房地产	房地产（0.034）、房地产楼盘展示（0.028）、房地产三维仿真展示（0.027）……
		旅游	文化旅游（0.075）、旅游系统（0.041）、模拟旅游（0.035）、智能旅游（0.033）、智能导游（0.023）……
		工业	工业设备（0.021）、工业自动化（0.015）、工业制造（0.014）、喷涂工业机器人（0.013）……
		军事	虚拟军事训练（0.033）、军事演习（0.033）、仿真枪（0.032）、仿真飞机驾驶舱（0.028）……
		医疗	医疗手术模拟（0.072）、医疗教学（0.054）、远程医疗（0.054）、医疗辅助（0.042）……
		教育	教育模拟（0.134）、预测教育（0.077）、远程教育（0.053）、安全教育（0.047）、智慧教育（0.042）……
		娱乐	互动娱乐（0.065）、健身（0.046）、虚拟朋友（0.042）、飞行影院（0.035）……

表 6 VR 产业链各环节技术主题的强度分布

产业链	主题编号	主题	LDA权重*专利数量权重	综合主题强度
上游-工具/设备设计	Topic1	输入设备	0.339 469 444	0.242 395 259
	Topic2	显示设备	0.123 443 434	0.088 143 730
	Topic3	拍摄设备	0.241 743 392	0.172 614 805
	Topic4	软件	0.419 193 329	0.299 321 418
中游-内容制作	Topic5	影视	0.012 320 067	0.008 797 039
	Topic6	声音	0.006 726 474	0.004 802 981
	Topic7	游戏	0.008 956 831	0.006 395 548
下游-行业应用	Topic8	房地产	0.028 126 076	0.020 083 185
	Topic9	旅游	0.039 397 977	0.028 131 789
	Topic10	工业	0.015 136 552	0.010 808 126
	Topic11	军事	0.030 380 456	0.021 692 906
	Topic12	医疗	0.040 149 437	0.028 668 363
	Topic13	教育	0.053 890 421	0.038 479 995
	Topic14	娱乐	0.041 545 006	0.029 664 857

知识管理论坛，2020(3):135-146
DOI：10.13266/j.issn.2095-5472.2020.013

示。由表 6 的综合主题强度可知，上游“工具 / 设备设计”产业链的强度最大，是当前最热门的研究领域；其次是下游“行业应用”，也是中国 VR 领域研发的共同关注焦点；最后是中游“内容制作”产业链，是中国 VR 领域研发的薄弱环节。从表 6 所示的各技术主题强度分布来看，在“工具 / 设备设计”环节，研发热点集中在 Topic4 软件研发和 Topic1 输入设备；在“内容制作”环节，研发热点集中在 Topic5 影视；在“行业应用”环节，研发热点集中在 Topic12 医疗、Topic13 教育、Topic14 娱乐。

5.4 产业链视角下的技术发展趋势分析

基于产业链语料库中的 IPC 编码，统计分析了中国 VR 专利的热点技术领域，部分统计结果如表 7 所示。结合表 7 的专利热点技术领域以及表 6 的 VR 产业链各环节技术主题的强度分布，可以分析出未来 5-10 年中国 VR 专利的发展趋势。具体体现在：①中国 VR 专利研发在产业链各个环节均会呈上升趋势，且上游研发与中、下游研发之间的增长幅度会渐渐趋于一致，三者之间的专利数量差距会缓慢减小。②热点研发环节仍会集中在上游的“工具 / 设备设计”，且以 G06（计算；推算；计数）、G02（光学）、H04（电通信技术）等技术领域为主流。③产业链下游的“行业应用”研发环节将会呈迅猛增长态势，且以 A63（运动；游戏；娱乐活动）、A61（医学或兽医学；卫生学）、E04（建筑物）等技术领域为主流。④产业链中游的“内容制作”作为薄弱研发环节会保持缓慢上升，且以 B60（一般车辆）、F16（工程元件或部件；为产生和保持机器或设备的有效运行的一般措施）、B64（飞行器；航空；宇宙航行）等技术领域为主流。

表 7 中国 VR 专利热点技术领域

IPC大类	技术领域含义	对应产业链	专利数量/件	比例/%
G06	计算；推算；计数	上游	3 991	38.61
G02	光学	上游	1 720	16.64
H04	电通信技术	上游	1 129	10.92
A63	运动；游戏；娱乐活动	下游	675	6.53
A61	医学或兽医学；卫生学	下游	319	3.08
E04	建筑物	下游	74	0.02
B60	一般车辆	中游	120	1.16
F16	工程元件或部件；为产生和保持机器或设备的有效运行的一般措施	中游	103	0.99
B64	飞行器；航空；宇宙航行	中游	77	0.03

6 结论与展望

6.1 研究结论

研究主要得出以下几个方面的结论：

（1）在产业链的构建与验证方面，结合 IPC 分类号、K-means 聚类的定性和定量分析可知，中国 VR 专利的上游材料端为“工具 / 设备设计”，中游生产端为“内容制作”，下游应用端为“行业应用”；且上游材料端专利聚类数量＞下游应用端专利聚类数量＞中游生产端专利聚类数量。此外，不仅在专利数量方面，而且在专利文本挖掘方面，目前中国 VR 行业更加注重上游产业端专利，且上游专利和中、下游专利之间的差距较大，产业链各环节的专利比例不均衡。

（2）在研发主题分布方面，结合 VR 产业链语料库的 TF-IDF 关键词权重值、LDA 概率权重值可知，“发明专利”类型在产业链上、



中、下游的比重均较大。此外，上游研发主题包括输入设备、显示设备、拍摄设备、软件等工具 / 设备，其中“发明专利”类型占比较大；中游研发主题包括影视、声音、游戏等内容制作，其中“外观设计”类型占比较大；下游研发主题包括房地产、旅游、工业、军事、医疗、教育、娱乐等行业应用，其中“实用新型”类型占比较大。

（3）在主题强度与研发热点挖掘方面，结合产业链视角下各主题的研发强度可知：综合主题强度中，上游是当前最热门的研究链，其次是下游产业链，最薄弱的是中游产业链，这与 IPC 分类号、K-means 聚类结果相一致，进一步验证了研究结果的科学性。此外，从各技术主题的强度分布来看，上游研发热点为输入设备和软件，诸如“信息输入设备”“数据输入设备”“客户端输入设备”等；中游研发热点为影视，诸如“VR 高清立体影视柔性传输线”“用于虚拟现实影视制作的稳拍系统”“VR 影视拍摄履带车”“用于播放 3D 影视的 VR 眼镜”等；下游研发热点为医疗、教育、娱乐，医疗诸如“基于 VR 技术的医疗手术模拟仿真系统”“基于虚拟现实的医疗设备操控系统”“基于虚拟现实的医疗设备演示系统”等，教育诸如“基于 VR 技术的小学生科技教育系统”“基于 VR 和动作捕捉的远程教育系统”“VR 安全教育动感座椅”等，娱乐诸如“三自由度虚拟现实游乐设备”“基于真实球拍的协同式增强现实乒乓球系统”“虚拟与现实有机结合的开心农场及实现方法”等。

（4）在技术发展趋势方面，中国 VR 专利研发在产业链各个环节均会呈上升趋势，且上游研发与中、下游研发之间的专利差距会缓慢减小。通过进一步细分的 IPC 分类号可知，产业链上游“工具 / 设备设计”的具体技术研发趋势为 G06F（计算；推算；计数——电数字数据处理）、G02B（光学——光学元件、系统或仪器）、H04N（电通信技术——图像通信，如电视）等领域；产业链下游“行业应用”的具体技术研发趋势为 A63F（运动；游戏；娱乐活动——利用小型运动物体的室内游戏）、A61B（医学或兽医学；卫生学——诊断；外科；鉴定）、E04H（建筑物——专门用途的建筑物或类似的构筑物）等领域；产业链中游“内容制作”的具体技术研发趋势为 B60R（一般车辆——不包含在其他类目中的车辆、车辆配件或车辆部件）、B64D（飞行器；航空；宇宙航行——用于与飞机配合或装到飞机上的设备；飞行衣；降落伞；动力装置或推进传动装置的配置或安装）、F16F（工程元件或部件——弹簧；减震器；减振装置）等领域。

（5）研究虽然是以 VR 专利领域为例进行实证分析，但相关研究思路、研究框架和研究方法可扩展到其他领域进行专利分析应用。在数据采集与清洗的基础上，基于 IPC 分类号构建专利产业链，并通过 K-means 聚类进行产业链验证，通过定性和定量研究方法的结合，而不仅仅是通过单一的定性方法，进行专利产业链的构建，为专利领域产业链研究提供新的研究视角。在产业链视角下，通过计算上、中、下游专利文本的关键词权重、主题词权重，进而结合二者衡量专利的综合强度，以此挖掘专利的技术主题强度与热点，预测专利的技术发展趋势，为专利文本挖掘和技术主题分析提供新的研究思路。

6.2 对策建议

中国 VR 正处于产业爆发的前夕，即将进入持续高速发展的窗口期。可以预见，在未来的五年内，VR 消费市场将迅速爆发，行业应用有望全面展开，文化内容将日趋繁荣，技术体系和产业格局也将初步形成。为推动我国 VR 产业发展，建议从以下方面开展工作：①进一步加强虚拟现实技术的研发。政府应支持设立重大相关研发项目，为产业发展提供共性技术、关键技术甚至颠覆性技术的供给；围绕虚拟现实产业链的关键环节，加强产学研合作，积极引导企业与科研单位投入虚拟现实研究，在关键技术上开展深度合作。②大力促进虚拟现实技术的市场化和产业化。以虚拟现实技术在工

ChinaXiv合作期刊
知识管理论坛，2020(3):135–146
DOI：10.13266/j.issn.2095–5472.2020.013

业、文化、教育、娱乐和医疗等领域带来的广阔前景为契机，明确产业政策支持的方向。③尽快建立虚拟现实技术的行业标准。形成我国虚拟现实技术标准体系，巩固自主技术布局占位，提高产业自主话语权。

6.3 研究展望

研究的局限性在于选取的检索数据库为“中国专利数据库”，数据仅限于在华申请的专利，且数据库没有相应的引文数据，无法做到与引文指标的对比分析。因此，在下一阶段的研究中，可以选择德温特专利数据库（Derwent Innovation Index，DII）作为检索数据库，德温特数据库及其专利引文索引涵盖100多个国家、40多个专利机构，数据最早可追溯至1963年，为大规模的专利文献研究提供了规范可靠的数据来源，而它的及时更新又为专利技术前沿的研究提供了可能，是企业和相关研究人员分析专利情报必不可少的工具。

参考文献：

[1] 杨超 , 朱东华 , 汪雪锋 , 等 . 专利技术主题分析 : 基于 SAO 结构的 LDA 主题模型方法 [J]. 图书情报工作 , 2017, 61(3): 86-96.

[2] 李姝影 , 张鑫 , 许轶 , 等 . 核心专利集筛选及专利技术主题识别影响 [J]. 情报学报 , 2019, 38(1): 17-24.

[3] 陈伟 , 林超然 , 李金秋 , 等 . 基于 LDA-HMM 的专利技术主题演化趋势分析——以船用柴油机技术为例 [J]. 情报学报 , 2018(7): 732-741.

[4] 张婷婷 . 网络综合布线实验室虚拟现实技术下的设计与研究 [J]. 电子测试 , 2019(3): 106-107.

[5] 孙柏林 . 区块链 + 虚拟技术 : 仿真技术的新动向 [J]. 计算机仿真 , 2019, 36(1): 8-13, 35.

[6] 周永伟 . 岩石工程虚拟现实系统的建立及应用 [J]. 山西建筑 , 2019, 45(2): 77-79.

[7] 谢敬伟 . 分布式虚拟现实交互仿真系统研究 [D]. 杭州 : 浙江大学 , 2017.

[8] PAN X, HAMILTON A F D C. Why and how to use virtual reality to study human social interaction: the challenges of exploring a new research landscape[J]. British journal of psychology, 2018:395-417.

[9] 刘子腾 . 面向虚拟人体解剖模型的交互式立体显示方法研究 [D]. 哈尔滨 : 哈尔滨工业大学 , 2017.

[10] 李发达 . 基于多传感器的交通控制硬件在环仿真技术研究与应用 [D]. 北京 : 北京工业大学 , 2017.

[11] 周雪 , 李飒 . 基于真实感图形生成技术的三维偶动画创作探索 [J]. 中国教育技术装备 , 2017(16): 43-44.

[12] 许兵 . 基于虚拟现实设备的典型飞机机翼装配仿真 [D]. 沈阳 : 沈阳航空航天大学 , 2017.

[13] 阎丽 , 胡丹丹 , 阎春元 , 等 . 基于感知觉学习的儿童视觉及智能虚拟现实数据库系统对弱视治疗效果的研究 [J]. 临床医学工程 , 2006(2): 32-33.

[14] 周哲泓 , 薛锦云 , 黄捷文 . 虚拟现实软件系统开发方法研究 [J]. 计算机工程与科学 , 2019, 41 (11): 1968-1975.

[15] 申闫春 , 王锐 , 郭富荣 , 等 . 基于并行渲染的虚拟现实开发平台设计与实现 [J]. 计算机仿真 , 2012, 29(11): 24-27.

[16] 刘园 .VR 技术在教育领域的研究与应用 [J]. 电脑知识与技术 , 2016, 12(16): 207-208.

[17] BAÑOS R M, GUILLEN V, QUERO S, et al. A virtual reality system for the treatment of stress-related disorders: a preliminary analysis of efficacy compared to a standard cognitive behavioral program[J]. International journal of human-computer studies, 2011, 69(9): 602-613.

[18] 陆颖隽 , 程磊 . 基于虚拟现实技术的图书馆信息资源建设与服务创新研究——以 CADAL 为例 [J]. 图书与情报 , 2017(4): 8-12.

[19] 丁铮 . 增强现实和虚拟现实在博物馆的应用 [J]. 信息与电脑 (理论版), 2017(24): 47-50.

[20] 胡阿沛 , 张静 , 雷孝平 , 等 . 基于文本挖掘的专利技术主题分析研究综述 [J]. 情报杂志 , 2013(12): 88-92.

[21] 张彬 , 陈永翀 , 张艳萍 , 等 . 锂浆料电池国际专利技术分析 [J]. 储能科学与技术 , 2017(5): 1000-1007.

[22] 隗玲 , 许海云 , 刘春江 , 等 . 技术领域主题发现研究——以基因工程疫苗领域为例 [J]. 数字图书馆论坛 , 2017(1): 39-47.

[23] 张杰 , 刘美佳 , 翟东升 . 基于专利共词分析的 RFID 领域技术主题研究 [J]. 科技管理研究 , 2013, 33(10): 129-132, 140.

[24] 林广杰 . 基于频繁项集的海量文本聚类研究 [D]. 北京 : 北京邮电大学 , 2015.

[25] 丁麒 , 庄志画 , 刘东丹 . 基于文本数据挖掘技术的 95598 业务工单主题分析应用 [J]. 电力需求侧管理 , 2016(A01): 55-57.

[26] 王镠富 , 胡等金 . 基于产业链的专利丛林测量与对策研究 [J]. 情报理论与实践 , 2019, 42(4): 101-106.

[27] 张杰 , 赵君博 , 翟东升 , 等 . 基于主题模型的微藻生物燃料产业链专利技术分析 [J]. 数据分析与知识发现 , 2019, 3(2): 52-64.

[28] 吴红 , 伊惠芳 , 马永新 , 等 . 面向专利技术主题分析的

WI—LDA 模型研究 [J]. 图书情报工作 , 2018(17): 68-74.
[29] 张超 . 基于专利数据挖掘的技术趋势分析方法 [D]. 大连 : 大连理工大学 , 2014.
[30] 林志坚 , 谌凯 , 潘婷婷 , 等 . 国内外虚拟现实技术专利分析研究 [J]. 竞争情报 , 2018:24-32.
[31] 黄立业 , 赵辉 , 王坚 , 等 . 基于专利分析的产业竞争情报分析框架研究 [J]. 情报科学 , 2015(4): 59-63.
[32] 王静宇 , 刘颖琦 , KOKKO A . 基于专利信息的中国新能源汽车产业技术创新研究 [J]. 情报杂志 , 2016, 35(1): 36-42.
[33] 于申 , 杨振磊 . 全球人工智能产业链创新发展态势研究 [J]. 天津经济 , 2019(5): 13-18.
[34] 方思 , 李国秋 . 全球无人驾驶汽车专利分析——从产业链和技术链的二维角度 [J]. 竞争情报 , 2016,12(5): 27-36.

作者贡献说明：
陈　玲：数据收集与分析，论文起草与修改，论文最终版本修订；
林　平：数据分析；
段尧清：提出整体研究思路与框架，修改论文。

Technology Topic Mining and Trend Analysis from the Perspective of Industrial Chain Combined with K-Means and LDA
——Taking Virtual Reality Technology as an Example

Chen Ling[1]　Lin Ping[1]　Duan Yaoqing[1,2]
[1]School of Information Management, Central China Normal University, Wuhan 430079
[2]Hubei Research Center of Data Governance and Intelligent Decision-making, Wuhan 430079

Abstract: [Purpose/significance] From the perspective of industry chain, this paper takes virtual reality technology as an example, constructs VR patent industry chain corpus, and explores the technical theme, research and development hotspot and future development trend of China VR patent. **[Method/process]** First of all, this paper used Python to crawl the patent text in VR field and got effective corpus through data cleaning. Secondly, combining IPC classification number and K-means clustering algorithm, this paper constructed and validates VR patent industry chain. In addition, based on TF-IDF algorithm and LDA theme model, we identified the core technology themes and their comprehensive strength, technology research and development hotspots and future trends of China VR patents from the perspective of production chain. **[Result/conclusion]** At present, the proportion of patents in each link of China VR industry chain is unbalanced. The upstream link is the most popular, followed by the downstream link, and the weakest link is the midstream link. In terms of theme mining, the upstream hot spot is software development, the midstream hot spot is film and television production, and the downstream hot spot is medical, educational and entertainment applications. In terms of future trends, the upstream of the industrial chain will be dominated by technologies such as electronic digital data processing, optical components, image communication, etc., the midstream will be dominated by technologies such as vehicle components, power devices, damping devices, etc., and the downstream will be dominated by technologies such as indoor games, medical diagnosis, identification, etc..

Keywords: K-means clustering algorithm　LDA theme model　technology theme evolution　text mining　VR